Tymen Preferred Flexible Analytic Process Guide

My Own Personal Preference Analytic Workflow

General workflow

1. Computing environment setup
2. Exploratory data analysis
3. A quick benchmark run
4. Data preprocessing
5. Feature enginnering
6. Feature selection
7. Model evaluation and selection
8. Parameter turning
9. Model ensembling
10. Prediction and submission

Computing environment setup

1. Use Google Cloud or Amazon AWS as computing platform

Exploratory data analysis

1. Calculate summary statistics.
   -. total number of samples and variables
   -. number of missing values and zeros
   -. mean, sd, min, max values for continuous variables
   -. number of unqiue values/categories for categorial and ordinal variables

2. Plot

A quick benchmark run

1. Use Random Forst (100 trees) without any feature enginnering to generate a quick submission. This submission can be used as a benchmark for further improvement. Plot the importance of the feaures to get a sense what are the most important features for prediction.
2. Train a simple Random Forest model and plot the confusion matrix for classfication or true-prediction-value-scatter-plot for regression. Find out where most the prediction errors come from. For example, it may come from certain categories. Need to split original training data into training and testing data.

Feature enginnering

1. General transformation: multiply, divide, sum, subtract, log, min, max, mean, std
2. If data have distance or length variables, several new features can be generated by multiplying (area or volume), dividing (ratio between two length), substrating (difference between two length) or summming (total length or distance) those variables or a subset of those variables.
3. Date variable:

(1) Extract day, month, quarter, year, weekend, weekday, holiday etc. as new features
(2) Calculate the length between two dates

Feature selection

1. Use the feature importance generated by Random Forest or XGBoost to rank features. Iteratively remove the least importance features and fit the model until the accuracy of the prediction to decrease.
2. Use XGBoost for feature selection:

(1) Keep the number of trees small (<20 trees)
(2) Keep the max depth of the tree small (<7)
(3) Iteratively run the feature importance analysis by removing the most(or least) important features.

3. Decision trees (XGBoost, Random Forest) are not affected multi-collinearity.

Model evaluation and selection

1. Popular models: Random Forst, Extra Trees, XGBoost

Parameter turning

1. Use grid search to fine tune parameters.
2. For Random Forest and Extra Trees, two important parameters that can be tuned: number of trees and the number of randomly selected features to seek split.
3. XGBoost:

(1) small eta -> small shinkage -> less overfitting -> slow convergence -> need more trees